Scalable Video Adaptation with Improved Perceptual Quality

Daxing Qian*, Hongyu Wang*, Zhan Ma†, and Wenzhu Sun*
* Dalian University of Technology. DaLian
† Dallas Technology Lab, Samsung Telecommunications America, Richardson, TX 75023, USA.
E-mail: qiandaxing23@sina.com, whyu@dlut.edu.cn, zhan.ma@ieee.org, sunwenzhu@mail.dlut.edu.cn

***Abstract*—Scalable video is an efficient and effective solution to stream video over heterogeneous networks to different clients, where a full resolution video bit stream can be adapted or truncated to meet the diverse network bandwidth requirements. Usually, there are several options to extract video streams for a given network bandwidth. To guarantee the high quality of service (QoS) or quality of experience (QoE) (for example, motion smoothness, etc) at a receiver, we propose an Equivalent Mean Square Error (Eq-MSE) scheme which is developed based on spatial and temporal frequency analysis of input video content. It is known that motion intensive content (like sports video) typically requires higher frame rate for smooth playback, and a relative lower frame rate is sufficient to provide decent visual quality for stationary videos. Proposed Eq-MSE is used to derive such minimal frame rate (MinFR) for different videos to guarantee motion smoothness. A simplified rate model is further introduced to obtain the quantization parameter (QP) given the MinFR, model parameters and network bandwidth. Thus, model derived QP and MinFR are employed to extract the proper video sub-stream from a full resolution scalable stream. It is noted that our proposed model based scalable adaptation is video content dependent. Compared with the default scalable video adaptation without considering the video content impact, our proposed scheme can provide better perceptual video quality by conducting the subjective video quality assessment.**

I. Introduction

With the advances of semi-conductor and access network technologies, real-time video streaming becomes more and more popular in our daily life. For example, we can easily enjoy the videos hosted at famous video-sharing communities (such as Youtube, Hulu, Tudou, etc) through wired or wireless networks using personal computer or mobile devices. How to provide the high quality of service (QoS) or quality of experience (QoE) to different users over heterogeneous networks is a crucial problem for the success of video streaming application. We propose to use the scalable video, where a full resolution scalable video stream can be adapted or truncated at the network gateway or proxy to meet different requirements imposed by the subscribed users and/or underlying access networks. We choose to use the scalable extension of the H.264/AVC (SVC) [1], [2] to enable the video bit stream scalability, due to its high coding efficiency and friendly network interface.

SVC includes temporal, spatial, SNR and combined scalabilities. Temporal scalability is enabled by the hierarchical prediction, such as hierarchical B-pictures [3]. Spatial scalability is achieved by encoding each supported spatial resolution into one layer. To improve the coding efficiency, inter-layer prediction is applied to remove the inter-layer redundancy, such as inter-layer intra prediction, inter-layer motion/mode prediction and residual prediction. SNR scalability includes coarse grain scalability (CGS) and medium grain scalability (MGS) [4]. CGS is a special example of the spatial salability with the same spatial resolution for different layers. To achieve the SNR refinement, we usually use different quantization parameters at different SNR layers. As an example, higher QP is chosen at lower SNR layer while finer QP is applied at higher SNR layer. It is noted that CGS provides limited bit stream extraction point (i.e., number of extraction point is the number of CGS layers). To provide more extraction points, MGS divides the refinement coefficients at enhancement layer into several fragments so that it can provide a progressive enhancement and graceful SNR degradation. On the other hand, MGS is not restricted to use the reference signal at current layer. Thus, coding efficiency is also improved for MGS by using reference pictures at enhancement layer. However, it also introduces the decoder drift if there is any packet loss at enhancement layer. Therefore, key picture [1] is used to reach the tradeoff between the coding efficiency and decoder drift. More information regarding the SVC techniques can be found in [1]. In this paper, we focus on the joint temporal and SNR scalability, and defer the spatial scalability as our future study.

As aforementioned, layered structure is employed in SVC to provide scalability. A typical full-resolution scalable video stream consists of a base layer (BL) and one or more enhancement layers (EL). Each enhancement layer is able to improve the resolution with respect to spatial, temporal or SNR. Fig. 1 depicts the layered SVC with one BL and two ELs. As shown, the smallest picture is reconstructed using the sub-stream extracted at base layer. It is QCIF (176 x 144) resolution with frame rate at 7.5 frames per second (fps). The intermediate picture is reconstructed using the base layer and enhancement layer #1, with CIF (352 x 288) resolution at 15 fps. The third picture is reconstructed by decoding all layers (i.e., 1 BL and 2 ELs). It is 4CIF (704 x 576) resolution at 30 fps. As known that BL can be decoded independently, while decoding EL requires the data of its reference layers. Different extraction points usually lead to different quality of experience. How to extract the proper sub-stream to meet the network bandwidth while providing enhanced QoE at receiver is a crucial problem for scalable video adaptation. In this work, we define the QoE as the perceptual video quality.

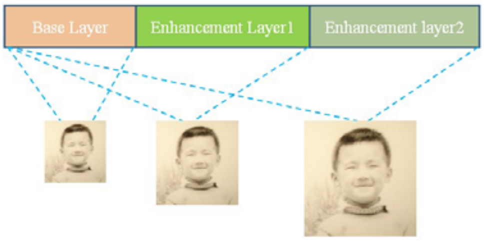


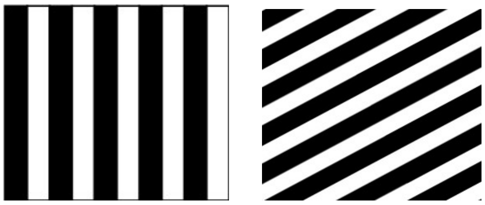
Fig. 1 SVC base layer and enhancement layers.

Usually, video content have a significant impact on the perceptual quality. For example, a sufficiently larger frame rate is necessary for a motion intensive video to maintain the continuity of the object movement to avoid jitter and guarantee the motion smoothness, while for stationary video, a relatively lower frame rate is enough to provide the decent video quality. For motion-intensive content, bit stream extracted at higher frame rate is favored. On the other hand, if there are larger high-frequency components (i.e., rich texture) in a single frame of the video, a finer quantization to reach better spatial quality is typically preferred. To solve this challenging problem, we analyze the spatial and temporal frequency of the input video content, and propose an Equivalent Mean Square Error (Eq-MSE) scheme to derive the minimal frame rate (MinFR) for different video sources to guarantee the motion smoothness and excellent QoE of the decoded video. We further simplify the rate model proposed in [9], and apply it to obtain the exact quantization parameter (QP) based on the network bandwidth requirement, MinFR and model parameters. In this work, we propose a simple table look-up method to estimate the model parameters. Alternatively, model parameters can be embedded in the full-resolution scalable video stream. It requires several bytes to embed the model parameters for a whole sequence which is far less than the video stream payload.

This paper is organized as follows. Section II introduces the temporal frequency induced by object movement in a video sequence (i.e., motion). In Section III we introduce spatial frequency of the general object in a picture and propose the Eq-MSE to derive the minimal frame rate for different input video sources. Simplified rate model is presented in Section IV together with the look-up table based model parameter prediction. Subjective test evaluation and experimental results are shown in Section V. Section VI concludes the paper and discusses the future directions.

II. TEMPORAL FREQUENCY INDUCED BY MOTION

Spatial frequency is introduced in [5]. We exemplify a simple case to show the spatial frequency in Figure 2.

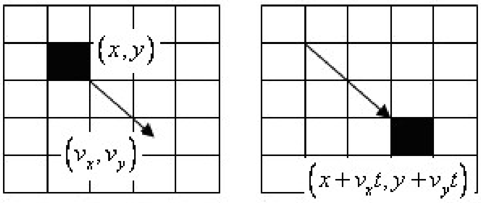


(a) (b)

Fig. 2 2D Sinusoidal signals.(a) $\left(f_x, f_y\right)$ = (5,0); (b) $\left(f_x, f_y\right)$ = (3,5).

Horizontal and vertical units are the width and height of the image, respectively. Therefore, $f_x = 5$ means that there are 5 cycles along each row.

Fig. 2(a) shows that the direction of spatial frequency is horizontal. If the plane moves vertically, then the eye will not perceive any changes no matter how fast the plane moves. Once its motion is tilted from the vertical direction, the eye will start to perceive temporal changes. The perceived change is most rapid when the plane moves horizontally.



(a) (b)

Fig. 3 Object is moving at velocity of $\left(v_x, v_y\right)$.

Fig. 3 depicts that every point $\left(x, y\right)$ at t = 0 is shifted by $\left(v_x t, v_y t\right)$ to $\left(x+v_x t, y+v_y t\right)$ at time t due to the motion. Alternatively, a point $\left(x, y\right)$ at time t corresponds to a point $\left(x-v_x t, y-v_y t\right)$ at time 0. Let the image of the object at time 0 be $\psi_0\left(x, y\right)$ and its moving velocities in horizontal and vertical directions by v_x and v_y. The image of the object at time t is:

$$\psi\left(x, y, t\right)=\psi_0\left(x-v_x t, y-v_y t\right) \qquad (1)$$

We perform the Continuous Space Fourier Transform (CSFT) on (1), where the CSFT of a signal $\psi(x)$ is defined as:

$$\Psi_c(f)=\int_{R^K}\psi(x)\exp(-j2\pi f^T x)dx \qquad (2)$$

where $f = [f_1, f_2, ..., f_K]^T \in R^K$ representing the continuous domain frequency variable.

Hence we can have:

$$\begin{aligned}
&\Psi\left(f_x, f_y, f_t\right)\\
&=\iiint \psi(x,y,t)\exp\left(-j2\pi\left(f_x x+f_y y+f_t t\right)\right)dxdydt\\
&=\iint \psi_0\left(x-v_x t, y-v_y t\right)\\
&\quad\cdot\exp\left(-j2\pi\left(f_x\left(x-v_x t\right)+f_y\left(y-v_y t\right)\right)\right)dxdy \qquad (3)\\
&\quad\cdot\int\exp\left(-j2\pi\left(f_t+f_x v_x+f_y v_y\right)t\right)dt\\
&=\Psi_0\left(f_x, f_y\right)\int\exp\left(-j2\pi\left(f_t+f_x v_x+f_y v_y\right)t\right)dt\\
&=\Psi_0\left(f_x, f_y\right)\delta\left(f_t+f_x v_x+f_y v_y\right)
\end{aligned}$$

where $\Psi_0\left(f_x, f_y\right)$ represents the 2D CSFT of $\psi_0\left(x, y\right)$. This function means that a spatial pattern characterized by $\left(f_x, f_y\right)$ in the object will lead to a temporal frequency, i.e.,

$$f_t = -f_x v_x - f_y v_y \qquad (4)$$

For a video signal, the temporal frequency is 2D position dependent. For a fixed 2D position (*x*, *y*), its temporal frequency is defined as the number of cycles per second usually denoted by Hertz (Hz).

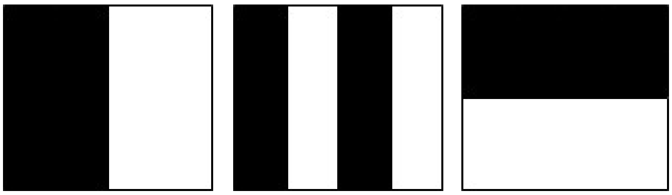
From (4) we can draw a conclusion that the temporal frequency depends on not only the motion, but also the spatial frequency of the object.

III. SPATIAL FREQUENCY OF GENERAL OBJECT IN A PICTURE

Video frame is typically represented by non-overlapped macroblock partitions for almost all video coding standards. SVC inherits the same MB partition from the H.264/AVC. Usually, an arbitrary object inside an image contains several macroblocks. In the following paragraphs, we introduce how to derive the SF of a general shape object inside an image or video frame.

A. The same f_t and the different MSE

We study the two special instances: the SF is $\left(f_x, f_y\right)$=(1,0) and $\left(f_x, f_y\right)$=(2,0), see from Fig. 4(a), (b). The size of a picture is $W \times H$. The objects of this two pictures are moving at the same speed: $\left(v_x, v_y\right)$=(*v*,0), we denote the value at (x, y) in picture by $f_t\left(x, y\right)$. After a sufficient short time Δt, the value is changed to $f_{t+\Delta t}\left(x, y\right)$.



(a) (b) (c)

Fig. 4 2D Sinusoidal signals.

(a) $\left(f_x, f_y\right)$ = (1, 0); (b) $\left(f_x, f_y\right)$ = (2, 0); (c) $\left(f_x, f_y\right)$ = (0, 1).

The equation of Mean Square Error (MSE) was introduced between the two pictures:

$$MSE\left(x, y\right) = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left[f_t\left(x, y\right) - f_{t+\Delta t}\left(x, y\right)\right]^2 \qquad (5)$$

The MSE of Fig. 4(a) at time *t* and $t+\Delta t$ is:

$$MSE_{4(a)}\left(x, y\right) = \frac{2}{WH}(H\times v\times\Delta t\times E\mathrm{rr})^2 = \frac{2H}{W}(v\times\Delta t\times E\mathrm{rr})^2$$

And the MSE of Fig .4(b) at time *t* and $t+\Delta t$ is:

$$MSE_{4(b)}\left(x, y\right) = \frac{4}{WH}(H\times v\times\Delta t\times E\mathrm{rr})^2 = \frac{4H}{W}(v\times\Delta t\times E\mathrm{rr})^2$$

Err is the difference gray value between black and white in Fig.4.

We can notice that the MSE value of Fig. 4(b) is twice larger than the same one of Fig. 4(a). If the SF is $\left(f_x, f_y\right)$ =(*n*,0), it will induce *n* times MSE value to $\left(f_x, f_y\right)$=(1,0).

Peak Signal Noise Ratio (PSNR) is commonly used to measure the coding efficiency, i.e.,

$$PSNR_{dB} = 10\lg^{\frac{\left(2^n-1\right)^2}{MSE}} \qquad (6)$$

where $\left(2^n-1\right)^2$ is the square of maximum possible signal value, *n* is the number of bits to represent each pixel. It is noted that the larger MSE is, the smaller PSNR is.

From Fig. 4(a) and Fig. 4(c), we can learn that the two pictures induce the same f_t but their MSE and PSNR are quite different. The utilization of MSE or PSNR as the adaptive parameter is not very reasonable. The Fig. 4(a) and Fig. 4(c) are moving at speed $(v,0)$ and $(0,v)$, respectively. As we can see, they have the same temporal frequency at $f_t = -v$, but their MSE values are obviously different with ratio at H/W .

B. Equivalent Mean Square Error (Eq-MSE)

We propose an Eq-MSE method to calculate the SF of general objects in a picture and find the appropriate frame rate [6], [7].

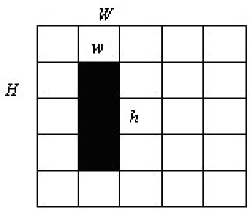


Fig. 5 Illustrative figure for object in general picture.

Fig. 5 illustrates that the size of black column is $w\times h$ and picture size is $W\times H$. We use f_{tB} to represent the induced frame rate by the object, which is defined as:

$$f_{tB} = -\frac{MSEf\left(\dfrac{h}{H}, 0\right)}{MSEf\left(1,0\right)}\cdot v_x - \frac{MSEf\left(0, \dfrac{w}{W}\right)}{MSEf\left(0,1\right)}\cdot v_y \qquad (7)$$

where $MSEf\left(f_x,0\right)$ is $MSE\left(x,y\right)$ when the picture SF is $\left(f_x,0\right)$, and $MSEf\left(0,f_y\right)$ is $MSE\left(x,y\right)$ when the picture SF is $\left(0,f_y\right)$. v_x and v_y are velocities in horizontal and vertical directions.

We regard the SF of a general picture is:

$$\left(f_x,f_y\right)=\left(\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf\left(1,0\right)},\frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf\left(0,1\right)}\right)=\left(\frac{h}{H},\frac{w}{W}\right) \quad (8)$$

The objects in a picture that induce the frame rate from moving from arbitrary directions are:

$$\begin{aligned} f_t &= \Sigma f_{tB} \\ &= \Sigma\left(-\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf\left(1,0\right)}\cdot v_x-\frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf\left(0,1\right)}\cdot v_y\right) \quad (9) \\ &= \Sigma\left(-\frac{h}{H}\cdot v_x-\frac{w}{W}\cdot v_y\right) \end{aligned}$$

where Σ is all the MBs in the picture. v_x and v_y are velocities in horizontal and vertical directions of corresponding MB. We get the mode and number of MB in a picture, and then choose the other picture within the same GOP to get MVs according to every MB. The ratio between MVs number in MB and the time interval between two frames are v_x and v_y. For example, the $\sum\frac{h}{H}v_x$ and $\sum\frac{w}{W}v_y$ of sequence Mobile are 12.2, 9.4, respectively. Its MinFR is 12.2+9.4=21.6. Note that with a real signal, the CSFT is symmetric, so that for every frequency component at $\left(f_x,f_y\right)$, there is also a component at $\left(-f_x,-f_y\right)$ with the same magnitude. The corresponding temporal frequency caused by this other component is $f_xv_x+f_yv_y$ [5].

Equation (9) is the function of minimal frame rate (MinFR) that makes the video motion smoothness without jitter.

IV. SIMPLIFIED RATE MODEL

During the past decade, lots of work has been done to make a significant improvement of SVC. As a kernel module, the extraction of bit stream was a topic of numerous research works. A major drawback of video coding method is that its prioritization policy is independent of the video content [8].

In [9], [10], Wang and Ma have proposed two analytical models regarding the rate and perceptual quality for scalable video focusing on the joint temporal and SNR scalability. Specifically, the rate model is the product of a power function of quantization stepsize q and frame rate t,

$$R(q,t)=R_{\max}(\frac{q}{q_{\min}})^{-a}(\frac{t}{t_{\max}})^{b} \quad (10)$$

where $R_{\max}=R(q_{\min},t_{\max})$, a and b are content dependent parameters, $q_{\min}$ and $t_{\max}$ are constants, equal to 16 and 30Hz, respectively. Different test sequences have different a values. They are quite similar. It is almost independent of video content. In this paper, we mainly analyze the video affected induced by frame rate.

As well known, within a GOP, the key frame is encoded independently while other frames refer to the key frame and the target of encoding is the difference between them. When the MVs are large in a video, there is significant difference between two frames. That means the data that will be encoded are large, and vice versa.

We use the equation

$$R(t;q)=\frac{R(q,t)}{R(q,t_{\max})} \quad (11)$$

to present normalized rate vs. temporal resolution (NRT) under the same quantization stepsize q. $R(q,t)$ is the bit rate obtained with chosen quantization stepsize q and frame rate t [10].

We consider two extreme situations. One situation implies that MVs are so large that it is entirely different between the frames in a GOP. Every frame needs to be encoded independently. The other one supposes no difference between two frames in a GOP. Only the key frame needs to be encoded. Fig. 6 depicts two NRT of extreme situations.

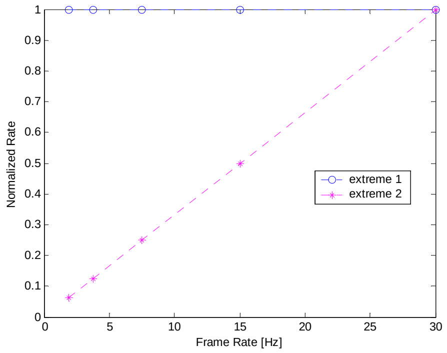


Fig. 6 Normalized rate vs. temporal resolution (NRT) of two extreme situations.

It is easy to understand that any curve of video sequences NRT is between the two lines. Due to the motion estimation, motion compensation techniques, the NRT curve is not linear. According to [9], [10], we suppose that $y=x^n$, x is the NRT, $n\in(0,1)$. Parameter n depends on the content of video sequence. When the video sequence has larger MVs, the n is closer to 1.

Parameter b of (10) indicates how fast the rate drops when the frame rate decreases, with a larger b indicating a faster drop. The higher motion sequences should have larger parameter b while lower motion sequences should have smaller one. The larger MVs are between two frames, the larger frame rate and parameter b will be, and vice versa.

We set parameter *b* to several discrete value $b \in \{0.06, 0.15, 0.20, 0.51\}$. According to MinFR of four sequences, we can get the predicted parameter *b* from Tab.1.

TABLE. 1 THE MAP OF THE PARAMETER B AND MINFR

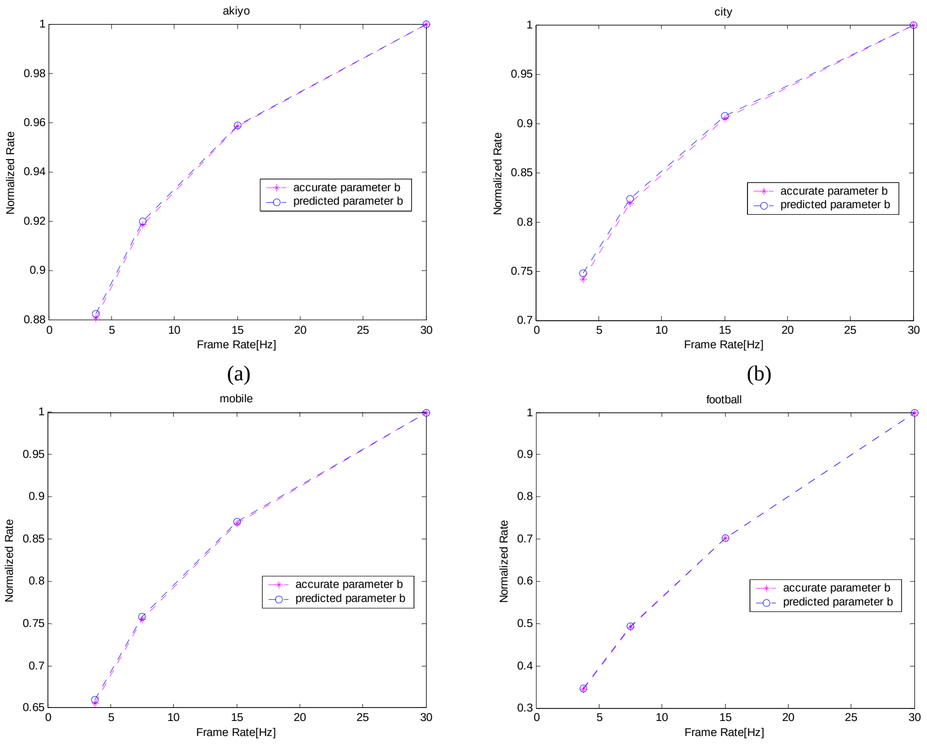
Parameter b	0.06	0.15	0.20	0.51
MinFR	[2,8)	[8,15)	[15,23)	[23,30]

Tab.2 depicts the parameters of different sequences for the rate model. We get the parameter *b* by using the least square method to make the least error of the rate model.

We use CIF test sequences to compare the NRT of model accuracy and the predict method which mentioned in this paper and use [11], [12] reference software code. Fig. 7 depicts the two curves of the NRT.

TABLE. 2 THE PARAMETERS FOR THE RATE MODEL

sequence	akiyo	city	mobile	football
a	1.021	1.057	1.007	1.064
b	0.061	0.144	0.203	0.512
predict b	0.06	0.14	0.20	0.51
RMSE/Rmax	2.10%	2.87%	2.07%	2.04%
Pre-b RMSE/Rmax	2.11%	2.88%	2.07%	2.04%



(a)

(b)

(c)

(d)

Fig. 7 Normalized rate vs. temporal resolution (NRT) using the accurate and predicted parameter b, respectively.

To the higher motion sequences, the NRT curve using accurate and predicted parameter *b* are almost coincide. While to the lower motion sequences, there is a little bit of difference at low frame rate between the two NRT curves. However, the error induced into the rate model could be ignored.

V. EXPERIMENTAL RESULTS

Due to 1.875 fps is rarely used, we set 8 pictures per GOP. Assuming maximum frame rate is 30 fps, we can have four different frame rates, i.e., 30, 15, 7.5 and 3.75 fps. Each video sequence is coded with 6 GOPs (48 frames) for simple.

The bit rate *R* is chosen as the extraction point of two sub streams. Its range should match extraction methods of sub0 and sub1. We extract the *Optimal Combination* (OC) sub stream *sub1* that is composed by MinFR and QP which is calculated by R and parameter *b*. Then we extract the *Frame Rate Different from OC* (FDOC) sub stream *sub0* that is constrained by bit rate *R* and frame rate *t*. Tab.3 depicts the target bit rate for each sequence. Moreover, the QP can be calculated by (10). Due to the parameter *a* has little difference among the four video sequences, we set $a = 1.02$ for simplicity. Finally, we extract sub stream *sub1* according to QP and MinFR.

TABLE. 3 THE MAXIMUM AND TARGET BIT RATE FOR SEQUENCES

Sequence	Akiyo	City	Mobile	Football
Max rate (kbps)	506	1296	2616	2917
Bit rate (kbps)	120	800	1000	1100

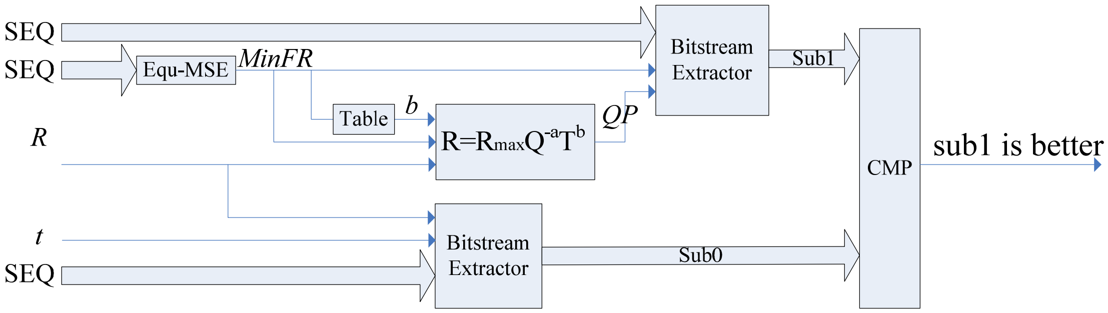


Fig. 8 Illustrative Flowchart for Bit Stream Extraction.

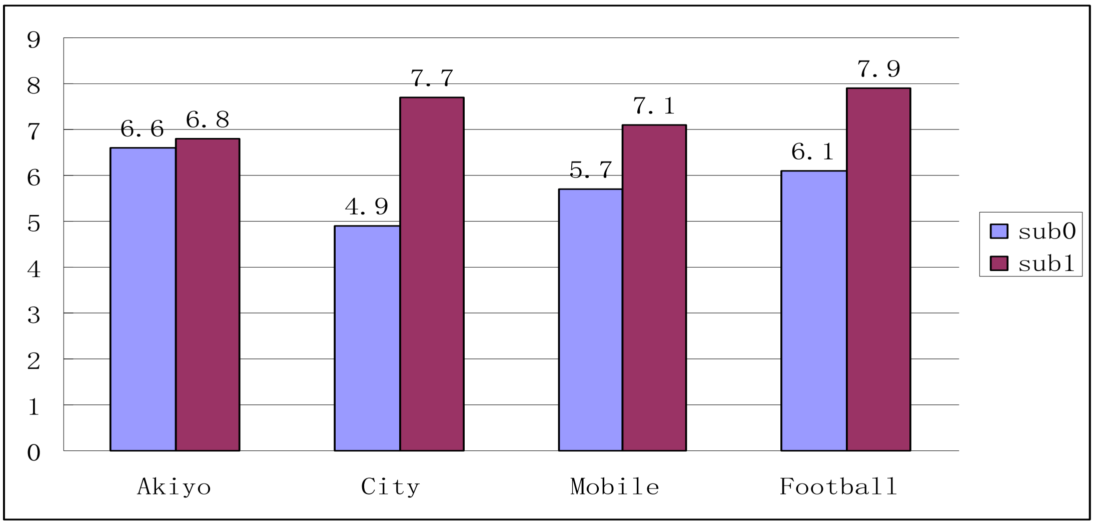


Fig. 9 Sequences subjective test comparative results of sub0 and sub1.

To evaluate the subjective quality of extracted video bitstream, we invite 15 viewers to give the subjective ratings for decoded video from FDOC sub stream sub0 and OC sub stream sub1 session, respectively. Sub0 is the default scalable video adaptation without considering the video content impact, while sub1 is our proposed model based on scalable adaptation with dependent video content. We use 11 ranks (i.e., 0~10) for the subjective tests ranging from the worst to the best and conduct the subjective assessments strictly following [13]. Fig. 9 depicts the subjective test results of four sequences. It is noted that “City”， “Mobile” and “Football” apparently have better perceptual rating for sub1 session, while “Akiyo” is quite similar between sub1 and sub0. We can see that the Eq-MSE method is providing better-decoded video quality at a given bit rate.

VI. Conclusions

In this paper, we propose the Eq-MSE scheme, which is developed based on the spatial and temporal frequency analysis of the video content. This scheme is used to derive the MinFR for different videos and in consequence, so as to guarantee the motion smoothness for decent decoded video quality. A simplified rate model is further introduced to obtain the QP given the MinFR, model parameters and network bandwidth. Thus, model derived QP and MinFR are employed to extract the proper video sub-stream from a full resolution scalable stream. Our proposed model is based on scalable adaptation and video content dependent. Compared with the default scalable video adaptation without considering the video content impact, our proposed scheme can provide better perceptual video quality at the same bit rate according to the subjective quality assessments. The scheme is well suited for practical applications. In future research, we will further investigate more video contents to verify our proposed method.

References

[1] Text of ISO/IEC 14496-10:2005/FDAM 3 Scalable Video Coding, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG, Lausanne, N9197, Sep. 2007.

[2] ISO/IEC ITU-T Rec. H264: Advanced Video Coding for Generic Audiovisual Services, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG, Int. Standard, May 2003.

[3] H. Schwarz, D. Marpe, and T. Wiegand, in: Hierarchical B pictures. Joint Video Team, Doc. JVT-P014, July 2005.

[4] Heiko Schwarz, Detlev Marpe, Thomas Wiegand, in: Overview of the Scalable Video Coding Extension of the H.264/AVC Standard, IEEE Transactions on Circuits and Systems For Video Technology, Vol. 17, No. 9, September 2007.

[5] Yao Wang, Jorn Ostermann, Ya-Qin Zhang, in: Video Processing and Communications. (2001)”

[6] DaXing Qian, HongYu Wang, FangLin Niu, “Scalable Video Coding Bit Stream Extraction Based on Equivalent MSE

Method," Advanced Materials Research Volx,204-210(2011) pp 1728-1732.

[7] Daxing A. Qian, Hongyu B. Wang, Wenzhu C. Sun and Kaiyan D. Zhu, "Bit Stream Extraction Based on Video Content Method in the Scalable Extension of H.264/AVC", accepted by Journal of Software, Feb,2011.

[8] Ehsan Maani, Aggelos K. Katsaggelos, Optimized Bit Extraction Using Distortion Modeling in the Scalable Extension of H.264/AVC, IEEE Transactions on Image Processing, Vol. 18, No.9, September 2009.

[9] Y. Wang, Z. Ma, and Y.-F. Ou, "Modeling Rate and Perceptual Quality of Scalable Video as Functions of Quantization and Frame Rate and Its Application in Scalable Video Adaptation," in Proc. of PacketVideo, May 2009.

[10] Zhan Ma, Meng Xu, Kyeong Yang and Yao Wang, "Modeling of rate and perceptual quality of video and its application to frame rate adaptive rate control", submitted to IEEE ICIP, Feb. 2010

[11] Joint Scalable Video Model JSVM 9_12_2 .

[12] JSVM Software Manual. JSVM 9.12.2 (CVS tag: JSVM_9_12_2) ,April 25th, 2008.

[13] ITU-R Rec. BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.